Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

An Approach to Accent Visualisation for the Reduction of Vowel Pronunciation Errors

Tom A. F. ANDERSON[a*], Barry Lee REYNOLDS[b] & David M. W. POWERS[a]
[a] *Centre for Knowledge and Interaction Technology, Flinders University, Australia*
[b] *Faculty of Education, University of Macau, Macau SAR, China*
*tom.anderson@flinders.edu.au

Abstract: In this paper, we introduce a novel mechanism for pronunciation evaluation for use in a computer-based application for the reduction of vowel errors. Unlike modern pronunciation feedback, which relies on the charts used by linguists or simply highlighting of incorrect segments, our two-dimensional maps situate the utterances produced by the learner in the context of the phonemes of the target accent.

Keywords: Pronunciation, phonemes, technology-enhanced language learning

1. Introduction

The concept behind this study is towards enhancing the activity of pronunciation practice. Learners know that they should devote time to pronunciation practice; however, efforts to substantially improve pronunciation require not only many sessions of practice but also integrated feedback. Learners need to understand how their pronunciation is improving, and the speaking quality needs to be reflexive. This means that learners should be able to practice the same thing many times with feedback that encourages the learner to alter their behaviour. Such practice induces the frequency effect, whereby "the greater the practice, the greater the performance" (Ellis, 2012, p.7).

A human teacher can provide targeted feedback to a learner, but there would be many advantages to computer-based training, including cost and repeatability (Chan et al., 2006). In mobile apps for language learning, the aspect of pronunciation is often restricted to reading (both alphabetic and phonetic) and listening. Reading and listening are necessary components for language learning; however, allowing users the opportunity to produce speech and providing visualisations that aid their understanding of their pronunciation errors may enhance the model of the student's current communicative abilities.

2. Literature Review

2.1 Pronunciation

The typical feedback for pronunciation provided by language learning software has room for improvement. A common mechanism, adopted in software such as Rosetta Stone, a leading language learning software, is providing spectrographs, voice contours for visual inspection of the voice patterns (Witt, 2012), but learners are not experienced linguists, so may miss the point of what such representations provide. Feedback based on automatic speech recognition provides an indication of which words were not pronounced correctly, so this gives learners a better idea of what to focus on.

Most state-of-the-art pronunciation modules use automatic speech recognition with posit (ASR) (Golonka et al., 2012). Typically, these systems, such as *EnglishCentral* or *Spexx,* identify those sections of utterances that are less than ideal (see Witt, 2012 for a more comprehensive review of these programs and more). However, due to the nature of current pronunciation training systems, they often do not provide an indication of how the sounds of the language relate to each other, or how the current utterance may relate to other similar utterances. With an aim to improve computer-based pronunciation

instruction, in this paper we present a biofeedback method that helps the learner to visualise their accent and how it fits into the context of the phonemic inventory of the language they are learning.

In contrast to systems that do not embed phonemes in their context, we have developed a neural network approach for visualising pronunciation. A Kohonen neural phonetic typewriter (KNPT) learns to represent the sounds of speech on a two-dimensional map with similar sounds located near one another (Kohonen, 1988; Kohonen, 2013). The underlying principle of the self-organising map is that information is arranged topographically, such that neurons that are near one another represent similar information. This allows viewing phonemes in relation to one another, and if necessary, by using an ensemble of maps, narrower contexts can also be visualised.

A two-dimensional map display allows the learner to get comfortable with a new representation of their voice. Although training a KNPT system can take a significant amount of time, once the values have been learned, it is very quick to classify the different sounds in a stream of speech. The learner can say something and the display will show the current state of their voice, along with the trajectory of the recent path of their voice. As neurons corresponding to similar inputs are located on nearby regions, a trajectory will generally pass from one zone to another through a transition zone. By tuning the maps, representations of voices and accents in the context of an individual speaker or an accent group are generated.

2.2 Form Focused Instruction

In recent years, isolated form-focused instruction has received renewed interest and a welcome response by second language acquisition researchers (Spada & Lightbown, 2008). Isolated form-focused instruction for pronunciation errors could be helpful in deterring first language interference with second language pronunciation. In terms of the kinds of mistakes that occur in pronunciation, Witt (2012) differentiated between phonemic and prosodic errors, the former being our current focus. Phonemic mistakes may arise in two forms: (1) severe - a phoneme is replaced by another, omitted, or an extra phoneme is produced; or (2) accented - a phoneme is pronounced with an accent. Its sound is thus different than a native speaker would produce. Both types of errors may affect the intelligibility of the learner.

Usually during form-focused instruction on such pronunciation errors, the classroom teacher provides oral corrective feedback in the form of recasts (implicit feedback) or metalinguistic explanation (explicit feedback). Although both are frequent occurrences in the second language classroom, explicit feedback has been shown to have a greater effect on language acquisition (Ellis, Loewen, & Erlam, 2006). However, in the limited class time a teacher has with students, all pupils cannot be given corrective feedback on their oral language production.

The effects arising from limited time for pronunciation feedback are sometimes alleviated by pairing language learners with tutors outside the classroom, which can result in not only the learners perceiving themselves as having improved but also showing said improvement on formal assessments (Lynch & Maclean, 2003). Still, this is not always a practical option since it cannot be guaranteed that a more capable peer or tutor can be secured for every language learner.

Form-focused explicit feedback given by the computer is a probable solution to this conundrum. In addition, one of the benefits of oral corrective feedback provided by the computer is the reduction in the potential of affective damage that can occur when language learners receive feedback from a teacher in front of their classroom peers. Language learners have an emotional response to the feedback delivered by teachers and when this response induces anxiety, the potential for negative effects on language learning increases (Agudo & de Dios, 2013). The computer can make accessible the type of feedback needed by all language learners within a comfortable context and environment.

3. Overview

A mobile application is used for accent reduction. A system flowchart of the system is depicted in Figure 1. For directed learning, a prompt is selected for a learner based on the system learner model. Note that learners may be involved in selecting the order of prompts, or produce spontaneous speech.

Speech is evaluated in the context of the accent of the speaker, with a goal to provide feedback quickly, within 300 milliseconds. The learner model is updated as interaction increases with the system.

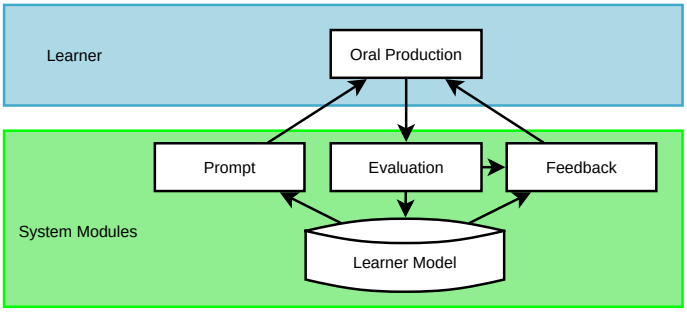

Figure 1. System Flowchart: Interactions between system modules and learner.

3.1 Improving the Pronunciation of Vowels.

As vowels are produced by an uninterrupted outflow of air, the sounds of vowels appear on the maps in continuous trajectories (averaged as traces with thickness reflecting variance). There are multiple ways to get similar sounds. However, the trace of a target accent for the pronunciation of the vowels is narrower than the trace for the corresponding foreign accented pronunciation. In other words, the sounds of the vowels that the foreign speakers produce are naturally and consistently out of the target zone. A speaker's first task is thus to pronounce the vowel closer to the target. The pronunciation of steady-state vowels can be changed and shaped by moving various articulators. The idea is to get the speaker to produce vowel sounds in isolation that are more similar to the target pronunciation. Next, speakers should produce the vowel sounds accurately in the context of isolated words.

In continuous speech, when many words are strung together, the pronunciation of each particular sound is much less important. In contrast, when a single word is pronounced in isolation, each word and each of its constituent phonemes is expected to be pronounced clearly and appropriately. The eventual goal is to aid the speaker to produce speech that is more intelligible or less accented. It is not guaranteed that reduction in accentedness in the pronunciation of isolated words will cross over to the regular speech patterns. However, during the process of learning how to make the sounds of an individual word be closer to the target pronunciation, the learner will gain an understanding of how to position their articulators to produce certain sounds that were previously less familiar to them. The next step would be to help learners to understand how the sounds of their own voice, in continuous speech, can be shaped to produce speech that is more like the target, and to give them an ability to practice shaping spoken words with less of an accent.

3.2 Speech Representation

Ensembles of self-organizing maps (SOMs) were trained on the voices of native speakers (general Australia, educated Melbourne) and a target group (Chinese background) using data from the AusTalk corpus (Burnham et al., 2011; Burnham et al., 2009; Wagner et al., 2010). The maps provide a visual representation of the speech of the learner in phonemic context. The speech of the learner is pre-processed into 39-feature mel-frequency cepstral coefficients (MFCCs), commonly used in automatic speech recognition. The MFCCs are then compared to codebook vectors. Although the initial training of the system must be performed offline, the evaluation and updates to the learner model can be performed near real-time. For more details, see Anderson and Powers (2016).

3.3 Implementation

Client – A microphone is used to obtain audio input. This may be the device microphone of the mobile computing device (often an array microphone in modern phones and laptops), but an external microphone may result in better sound quality.

System – As the learner speaks, their voice is analysed using the system and feedback is presented. Speech is shown on a map using a selection of pronunciation samples, as in Figure 2. Learners interactively explore the differences between how their speech and pre-recorded utterances are rendered, thereby improving the understanding of their speech in the context of the speech of others.

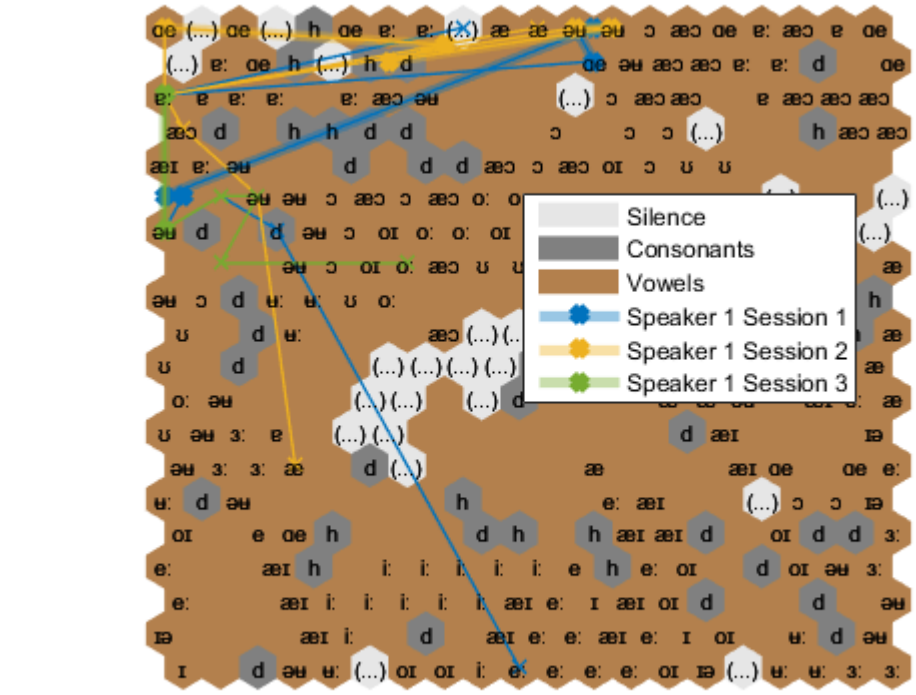

Figure 2. Trajectory analysis. The map depicts the trajectory of the learner's voice in the context of the different phonemes of the lesson (currently /ɐ/ as in "hard").

Gamification – Games can increase motivation. In a meta-analysis on digital game based learning for English as a foreign language, Chui et al. (2011) found that language learning games positively affect learning but meaningful and engaging games can yield a greater positive effect than drill-and-practice games. When learners are immersed in digital game based learning environments for more than a month, they gain automisation of language knowledge (Kao, 2014). We feel that automisation is a necessary aspect for improving pronunciation. In terms of future directions for our research, we aim to apply our SOM-based phoneme visualisation to game-based learning. The difficulty level can be increased by chaining multiple utterances together, providing fewer hints, or by narrowing the tolerance, which requires a learner to produce speech closer to the target.

4. Conclusion

In this paper, a mobile learning environment for pronunciation was proposed. Learners receive real time feedback about their speech and achieve goals in a video game by modifying their speech. Learners are presented with lessons sequenced so that they may gradually improve toward the target. In future research, the usability of the system for accent reduction would be evaluated, along with additional aspects for promoting learning such as gamification.

Acknowledgements

The AusTalk corpus was collected as part of the Big ASC project (Burnham et al. 2009; Wagner et al. 2010; Burnham et al. 2011), funded by the Australian Research Council (LE100100211). See: https://austalk.edu.au/ for details

References

Agudo, M., & de Dios, J. (2013). An investigation into how EFL learners emotionally respond to teachers' oral corrective feedback. *Colombian Applied Linguistics Journal*, *15*(2), 265-278.

Anderson, T. A. F., & Powers, D. M. W. (2016). *Characterisation of speech diversity using self-organising maps.* 16th Australasian International Conference on Speech Science and Technology (SST2016).

Burnham, D., Ambikairajah, E., Arciuli, J., Bennamoun, M., Best, C. T., Bird, S., Butcher, A. B., Cassidy, C., Chetty, G., Cox, F. M., Cutler, A., Dale, R., Epps, J. R., Fletcher, J. M., Göcke, R., Grayden, D. B., Hajek, J. T., Ingram, J. C., Ishihara, S., Kemp, N., Kinoshita, Y., Kuratate, T., Lewis, T. W., Loakes, D. E., Onslow, M., Powers, D. M., Rose, P., Togneri, R., Tran, D., & Wagner, M. (2009). A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus". In *Selected Proceedings of the HCSNet Workshop on Designing the Australian National Corpus* (pp. 96-107). Somerville, MA, USA: Cascadilla Proceedings Project.

Burnham, D., Estival, D., Fazio, S., Viethen, J. Cox, J., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow. M., Lewis, T., Butcher, A., & Hajek, J. (2011) Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box. In *Proceedings of 12th Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 841-844). France: International Speech Communication Association.

Chan, T., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., Patton, C., Cherniavsky, J., Pea, R., Norris, C., Soloway, E., Balacheff, N., Scardamalia, M., Dillenbourg, P., Looi, C., Milrad, M., & Hoppe, U. (2006). One-to-one technology-enhanced learning: An opportunity for global research collaboration. *Research and Practice in Technology-Enhanced Learning, 1*(1), 3-29.

Chiu, Y. H., Kao, C. W., & Reynolds, B. L. (2012). The relative effectiveness of digital game-based learning types in English as a foreign language setting: A meta-analysis. *British Journal of Educational Technology*, 43(4), 104–107.

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. T. G. D. Divjak (Ed.), *Frequency effects in language learning and processing* (pp. 7-33). Germany: Walter de Gruyten GmbH & Co.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in second language acquisition*, *28*(2), 339-368.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2012). Technologies for foreign language learning: A review of technology types and their effectiveness. *Computer Assisted Language Learning*, *27*(1), 70–105.

Kao, C. (2014). The Effects of Digital Game-based Learning Task in English as a Foreign Language Contexts: A Meta-analysis. *Education Journal,* 42(2), 113–141.

Kohonen, T. K. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65.

Kohonen, T. K., Torkkola, K., Shozakai, M., Kangas, J., & Venta, O. (1988). Phonetic typewriter for Finnish and Japanese. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing* (pp. 607–610). IEEE.

Lynch, T., & Maclean, J. (2003). Effects of Feedback on Performance: A Study of Advanced Learners on an ESP Speaking Course. *Edinburgh Working Papers in Applied Linguistics*, *12*, 19-44.

Spada, N., & Lightbown, P. M. (2008). Form-Focused Instruction: Isolated or Integrated? *Tesol Quarterly*, 42(2), 181-207.

Wagner, M., Tran, D., Togneri, R., Rose, P., Powers, D., Onslow, M., Loakes, D., Lewis, T., Kuratate, T., Kinoshita, Y., Kemp, N., Ishihara, S., Ingram, J., Hajek, J., Grayden, D. B., Göcke, R., Fletcher, J., Estival, D., Epps, J., Dale, R., Cutler, A., Cox, F., Chetty, G., Cassidy, S., Butcher, A., Burnham, D., Bird, S., Best, C., Bennamoun, M., Arciuli, J., & Ambikairajah, E. (2010). The Big Australian Speech Corpus (the Big Asc). In M. Tabain, J. Fletcher, D. Grayden, J. Hajek & A. Butcher (Eds), *13th Australasian International Conference on Speech Science and Technology* (pp.166-170). Melbourne: ASSTA.

Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, 1–8.